

机器学习理论与算法前沿讨论班

Syllabus

陆一平 北京大学北京国际数学研究中心

讨论班简介

本讨论班围绕机器学习中的泛化、生成、推理、后训练、规模定律与优化展开，关注实际现象背后的数学问题，以及理论如何帮助我们改进算法。通过文献阅读与专题讨论，希望大家形成对一个研究方向的整体认识，找到值得继续研究的问题。

机器学习理论的基础导读由陆一平主讲，后续希望每位同学围绕一个研究方向作一次专题报告。也欢迎同学提出希望陆一平介绍的研究，包括应用概率、随机模拟与采样、科学机器学习、推理侧算法和神经网络优化器等，具体专题根据大家的兴趣安排。

参与背景：建议具备概率论、线性代数、数学分析和机器学习基础，其他知识可随选题补充。讨论班以学生专题报告和共同讨论为主，时间、地点及报告顺序另行安排。

专题阅读分为扩散模型与语言模型两条主线。扩散模型部分从 Wasserstein 空间、梯度流与采样基础出发，讨论训练与泛化、离散建模和推理侧引导；语言模型部分从训练动力学与内部机制出发，依次讨论上下文学习与推理、推理侧扩展和后训练。机器学习理论作为共同基础，规模定律与优化器作为两条主线共享的专题。

选题与准备

下列论文表仅供选题参考，帮助大家了解研究问题与相关研究者。每位同学需要围绕自己的兴趣继续寻找文献，选择自己认为问题重要、质量高、值得深入理解的工作。选题和主讲论文**应提前与陆一平讨论**；阅读和准备报告的过程中，也请多交流问题、相关文献与新的想法。

报告应围绕一个明确的研究问题组织，介绍领域的发展脉络与不同工作的联系。建议精读两至三篇主论文，对照阅读四至五篇相关论文；较大的主题可以拆成连续几次报告。报告前分享论文链接与简短导读，并与陆一平确认范围。准备过程中，也可以请陆一平先介绍关键证明的基本思路。

准备报告时，建议围绕以下内容展开：

- 1. 问题与背景。**这个方向试图解释什么现象，解决什么问题，为什么值得研究？
- 2. 文献脉络。**已有方法分别依赖什么思路，主论文相对于相关工作推进了什么？
- 3. 主要结论。**说明问题设定、关键假设、误差指标及复杂度依赖；区分严格定理、近似分析和实验观察。
- 4. 核心方法。**讲清一个关键证明、算法构造或实验设计，让听众理解结论为何成立。
- 5. 局限与问题。**讨论假设是否合理、比较是否公平，以及一个可进一步验证或研究的具体问题。

报告顺序可先安排基础导读，再根据选题之间的联系展开。参考阅读中的年份主要按所链接 arXiv 的首次提交年份列出，会议或期刊年份可能不同；讲义与技术文章单独注明。

一 机器学习理论与泛化

阅读路线：从随机标签实验与惰性训练出发，联系核谱与计算最优规模定律，再深入平均场训练和高维 SGD 的有效动力学。

从逼近误差、统计误差与优化误差出发，理解过参数化模型的泛化。重点讨论 RKHS、核回归、神经切线核、特征学习，以及核谱和目标函数结构如何影响学习曲线与规模定律。基础导读由陆一平主讲，后续可选择以下工作继续深入。

统计学习基础：泛化理论、逼近理论与核学习，可参考陆一平的 [Statistical Learning 课程资料](#)。

1. Zhang、Bengio、Hardt、Recht、Vinyals (2016) . [Understanding deep learning requires rethinking generalization](#).

基础阅读：随机标签实验与有限样本表达能力，理解为什么仅依赖模型容量的传统解释不足以描述实际泛化。

2. Chizat、Oyallon、Bach (2018) . [On Lazy Training in Differentiable Programming](#).

理论主读：参数缩放如何使训练接近初始化处的线性化，以及核训练与特征学习之间的边界。

3. Elliot Paquette、Courtney Paquette、Xiao、Pennington (2024) . [4+3 Phases of Compute Optimal Neural Scaling Laws](#).

理论主读：在可解模型中，分析数据与目标复杂度如何决定计算最优的模型规模，以及容量、优化噪声和特征嵌入如何共同形成不同的规模区间。

4. Li、Chen、Huang、Wang、Wu (2025) . [Functional Scaling Laws in Kernel Regression: Loss Dynamics and Learning Rate Schedules](#).

阅读重点：如何用核模型刻画整个训练损失轨迹，以及学习率调度的影响。

5. Chizat、Bach (2018) . [On the Global Convergence of Gradient Descent for Overparameterized Models using Optimal Transport](#).

理论主读：将宽神经网络训练写成测度空间中的 Wasserstein 梯度流，理解初始化条件与多粒子极限如何支持全局收敛结论。

6. Ben Arous、Gheissari、Jagannath (2022) . [High dimensional limit theorems for SGD: Effective dynamics and critical scaling](#).

理论主读：通过低维统计量的 ODE 与 SDE 极限刻画高维 SGD，重点理解临界步长下的漂移修正如何改变有效动力学与学习相图。

7. Ren、Nichani、Wu、Lee (2025) . [Emergence and scaling laws in SGD learning of shallow neural networks](#).

阅读重点：如何由不同特征的学习时间尺度解释涌现现象与整体规模定律。

8. Bordelon、Canatar、Pehlevan (2020) . [Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks](#).

阅读重点：核谱及目标函数在谱方向上的分量如何共同决定学习曲线。

讨论问题：核理论能解释哪些现象？引入特征学习之后，样本复杂度、训练轨迹和规模定律可能怎样变化？

二 扩散模型

围绕扩散模型如何学习分布、如何生成样本以及如何适应目标任务展开。连续与离散模型可对照阅读，重点理解训练误差、数值误差和粒子近似误差分别如何影响生成结果。

Wasserstein 空间与梯度流基础

阅读路线：耦合与最优输运，Wasserstein 空间中的连续性方程与测地线，再到自由能、JKO 格式和熵耗散。先建立概率分布如何沿动力系统演化的数学语言，再用它分析采样算法。

基础内容：概率测度的推前与耦合、二阶 Wasserstein 距离、动态输运公式、测地凸性，以及自由能的一阶变分。以热方程和梯度型漂移的 Fokker Planck 方程为例，理解分布空间中的梯度流；结合 Poincaré、对数 Sobolev 与输运不等式，比较方差、相对熵和 Wasserstein 距离的收敛。

9. Santambrogio (2016) . [Euclidean, Metric, and Wasserstein Gradient Flows: an overview](#)。

基础导读：从欧氏空间最速下降进入度量空间与 Wasserstein 空间，学习最优输运、测地凸性、能量耗散与 JKO 格式；可作为本专题的入门主线。

10. Benamou、Brenier (2000) . [A computational fluid mechanics solution to the Monge Kantorovich mass transfer problem](#)。

几何基础：以连续性方程和速度场的动能表达二阶最优输运，把分布之间的距离转化为路径优化；理解概率密度、速度场与 Wasserstein 测地线的关系。

11. Jordan、Kinderlehrer、Otto (1998) . [The Variational Formulation of the Fokker Planck Equation](#)。

理论主读：以自由能和 Wasserstein 距离构造 JKO 变分离散格式，说明梯度型漂移下的 Fokker Planck 方程为何具有梯度流结构；重点理解离散能量估计与连续极限。

12. Otto、Villani (2000) . [Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality](#)。

进阶阅读：联系相对熵、Fisher 信息、Wasserstein 距离与输运不等式；关注对数 Sobolev 不等式如何控制熵耗散，并明确各个不等式成立的几何与正则性条件。

13. 苏剑林 (2023, 科学空间博客) . [梯度流：探索通向最小值之路](#)。

中文导读：从向量空间的最速下降进入 Wasserstein 梯度流，联系连续性方程与 Fokker Planck 方程；配合 Santambrogio 与 JKO 阅读。讨论几何的选择如何改变下降方向，以及测度演化如何对应粒子运动。

讨论问题：同一个分布演化能否有不同的梯度流解释？能量泛函与几何各自决定什么？JKO 每步的测度优化与粒子位置的数值更新有什么联系？连续时间的能量下降，如何转化为离散算法可用的误差界？

采样算法与复杂度基础

阅读路线：不变分布与混合时间，Langevin 扩散及其离散化，再到动量、Metropolis 校正、近端采样与 Stein 粒子方法。

基础内容：从只知道未归一化密度的采样问题出发，学习 Markov 核、不变分布、细致平衡与遍历性。比较总变差、KL 和 Wasserstein 误差，区分单个样本的分布误差与相关样本估计期望的方差。分析算法时明确初始化、步长、势函数正则性、泛函不等式常数，以及密度、梯度或采样 oracle 的调用成本。

Langevin 动力学与离散化

14. Sinho Chewi (2026 书稿) . [Log Concave Sampling](#)。

基础导读：先读第 1、2 章的随机微积分、Markov 半群、最优输运与泛函不等式，再选读离散 Langevin、Metropolis 校正和采样复杂度；按专题查阅，无需通读全书。

15. Dalalyan (2014) . [Theoretical guarantees for approximate sampling from smooth and log concave densities](#)。

采样理论主读：从连续 Langevin 扩散出发，建立离散 Langevin Monte Carlo 的非渐近误差界；理解混合误差、时间离散化偏差和初始化对维数及精度依赖的影响。

16. Wibisono (2018) . [Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem](#)。

几何与算法主读：把 Langevin 解释为测度空间中的复合优化，分析 ULA 的离散偏差与分裂方法；用高斯目标检验不同离散格式的平稳分布，并讨论近端步骤的可实现性。

17. Vempala、Wibisono (2019) . [Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices](#)。

理论主读：在目标满足对数 Sobolev 不等式、势函数 Hessian 有界的条件下分析 ULA 的 KL 误差，无需假设势函数凸；区分连续时间熵收敛与离散算法的偏差项。

讨论问题：以一维高斯目标为例，连续 Langevin 与固定步长 ULA 的平稳方差是否相同？混合时间与离散偏差如何共同决定最优步长？对数 Sobolev 条件比强对数凹性允许哪些分布，常数又如何影响实际复杂度？

动量 校正与近端采样

18. Cheng、Chatterji、Bartlett、Jordan (2017) . [Underdamped Langevin MCMC: A non asymptotic analysis](#)。

算法与理论主读：在光滑强对数凹条件下，引入速度变量并分析欠阻尼 Langevin 的收敛与离散化；比较它与过阻尼方法的维数、精度和条件数依赖。

19. Dwivedi、Yuansi Chen、Wainwright、Yu (2018) . [Log concave sampling: Metropolis Hastings algorithms are fast](#)。

校正方法主读：通过接受拒绝步骤使 Langevin 提议保留目标分布，分析 MALA 的非渐近混合时间；关注热启动、接受率与步长条件，以及对精度依赖的改善。

20. Yin Tat Lee、Ruoqi Shen、Kevin Tian (2020) . [Structured Logconcave Sampling with a Restricted Gaussian Oracle](#)。

进阶阅读：以受限高斯采样 **oracle** 构造近端式归约，处理复合目标并改善条件数依赖；明确一次 **oracle** 调用如何实现、代价多大，以及内层采样误差如何计入总复杂度。

讨论问题：加入速度、接受拒绝或高斯正则化，各自改变了哪一项复杂度？热启动成本应如何计入？有近似 **score** 时，哪些算法还需要额外的密度信息才能正确执行？

Stein 粒子方法与梯度流

通过相互作用粒子的输运近似目标分布，将测度空间优化与可计算的粒子更新联系起来。重点比较核诱导的几何、Langevin 的随机扰动，以及有限粒子带来的近似误差。

21. Qiang Liu、Dilin Wang (2016) . [Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm](#)。

粒子方法基础：从分布变换下 KL 的方向导数和 Stein 恒等式推导 SVGD，理解核平滑的 **score** 项与粒子排斥项，以及粒子数和核带宽对分布近似的作用。

22. Qiang Liu (2017) . [Stein Variational Gradient Descent as Gradient Flow](#)。

理论补充：研究 SVGD 的测度演化和 Stein 算子诱导的梯度流几何；区分这一几何与标准 Wasserstein 梯度流，并讨论连续分布极限、有限粒子近似和弱收敛条件。

讨论问题：连续分布层面的 KL 耗散能为有限粒子提供什么保证？粒子之间的相关性、核的选择与高维退化，如何影响样本质量和期望估计？

与扩散模型的衔接：前向热流和 Ornstein Uhlenbeck 过程提供梯度流与熵耗散的例子；反向生成则依赖时间反演和随时间变化的 **score**。接下来结合 Song 等的 SDE 框架与《Sampling is as easy as learning the score》，分析 **score** 估计误差、时间离散化、终端初始化和提前停止如何共同影响生成分布。

选题建议：基础报告可从 Santambrogio 与 JKO 入手；采样理论可围绕 Dalalyan 与 Vempala、Wibisono，或欠阻尼 Langevin 与 MALA 对照；进阶报告可选择近端采样或 SVGD。每次围绕一个问题和一篇主读论文展开，其余作为背景，具体选择与陆一平讨论确定。

训练理论与泛化

阅读路线：连续时间表述与扩散模型设计，**score** 误差到采样误差的传递，再到统计最优性、训练动力学和记忆泛化分离。

理解 **score matching** 如何把分布学习转化为函数估计，以及估计、优化、离散化和采样误差之间的关系。在此基础上讨论模型何时记忆训练样本、何时学习总体分布，重点比较网络容量、训练时间、噪声水平和采样方法的作用。

23. Song 等 (2020) . [Score Based Generative Modeling through Stochastic Differential Equations](#)。

基础阅读：正向与反向 SDE、**score estimation**，以及生成模型的连续时间表述。

- 24.** 苏剑林（2022，科学空间博客）. [生成扩散模型漫谈（五）：一般框架之 SDE 篇](#)。
中文导读：梳理前向加噪、反向 SDE 与得分匹配，作为 Song 等 SDE 论文的推导补充。报告时明确反向时间的约定，并解释反向漂移为什么需要随时间变化的 score。
- 25.** 苏剑林（2022，科学空间博客）. [生成扩散模型漫谈（六）：一般框架之 ODE 篇](#)。
中文导读：由 Fokker Planck 方程理解概率流 ODE，并联系 DDIM。讨论精确 score 下边缘分布一致所需的条件，以及这种一致性为何不意味着 SDE 与 ODE 的样本路径或离散误差相同。
- 26.** Karras、Aittala、Aila、Laine（2022）. [Elucidating the Design Space of Diffusion Based Generative Models](#)。
基础阅读：将噪声参数化、网络预条件、训练目标与采样器分开分析，理解不同设计选择如何影响生成质量与采样效率。
- 27.** 苏剑林（2023，科学空间博客）. [生成扩散模型漫谈（二十一）：中值定理加速 ODE 采样](#)。
中文导读：从 Euler、Heun 方法进入 AMED 的定制求解器，与 EDM 对照。比较相同网络调用次数下的误差与额外蒸馏成本；留意标量积分中值定理向向量场推广时所用的近似。
- 28.** Chen、Chewi、Li、Li、Salim、Zhang（2022）. [Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions](#)。
理论主读： L^2 score 误差如何传递为生成分布误差，以及数据矩、正则性和离散化在采样复杂度中的作用。
- 29.** 苏剑林（2023，科学空间博客）. [生成扩散模型漫谈（十六）：W 距离 \$\leq\$ 得分匹配](#)。
中文导读：理解扩散模型中得分匹配误差对 Wasserstein 生成误差的控制，衔接《Sampling is as easy as learning the score》。比较不同误差指标，并核对界中的权重、常数、终端分布项和成立条件。
- 30.** Oko、Akiyama、Suzuki（2023）. [Diffusion Models are Minimax Optimal Distribution Estimators](#)。
阅读重点：在给定函数类和经验目标优化条件下，扩散模型如何达到接近 minimax 的分布估计率。
- 31.** Biroli、Bonnaire、de Bortoli、Mézard（2024）. [Dynamical Regimes of Diffusion Models](#)。
阅读重点：反向生成过程中的不同动力学阶段，以及数据结构与记忆之间的关系。
- 32.** Bonnaire、Urfin、Biroli、Mézard（2025）. [Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training](#)。
阅读重点：泛化与记忆出现的不同训练时间尺度，以及训练动力学的隐式正则化作用。
- 33.** Farghly、Rebeschini、Deligiannidis、Doucet（2025）. [Implicit Regularisation in Diffusion Models: An Algorithm Dependent Generalisation Analysis](#)。
阅读重点：score stability 如何给出依赖算法的泛化分析，以及训练和采样中的正则化来源。

34. Ye、Zhu、Tao、Chen (2025) . [Provable Separations between Memorization and Generalization in Diffusion Models](#).

阅读重点：从统计估计和网络逼近两个角度，区分总体 score 与经验 score 的学习难度。

讨论问题：不同论文如何定义记忆与泛化？比较结论时，应区分训练迭代时间与反向生成时间，并说明所用距离及模型假设。

离散扩散模型

阅读路线：D3PM 与连续时间马尔可夫链，MDLM 与 SEDD 的目标函数，再到条件分布解释、高阶求解器、统计最优性与验证器引导采样。

从连续时间马尔可夫链、离散 score 和 score entropy 损失出发，理解离散扩散与条件分布估计、自回归建模的关系。进一步讨论统计估计的最优性，以及数值求解和并行生成的计算代价。

35. Austin、Johnson、Ho、Tarlow、van den Berg (2021) . [Structured Denoising Diffusion Models in Discrete State Spaces](#).

基础阅读：D3PM 的转移矩阵、吸收状态与变分目标，理解离散空间中如何设计加噪过程。

36. Campbell、Benton、De Bortoli、Rainforth、Deligiannidis、Doucet (2022) . [A Continuous Time Framework for Discrete Denoising Models](#).

理论主读：用连续时间马尔可夫链统一离散扩散，推导反向跳跃率、训练目标及采样分布的误差保证。

37. Sahoo 等 (2024) . [Simple and Effective Masked Diffusion Language Models](#).

基础阅读：MDLM 如何化简吸收扩散的训练目标，将其写成掩码语言建模损失的混合，并利用 Rao Blackwell 化降低估计方差。

38. Lou、Meng、Ermon (2024) . [Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution](#).

基础阅读：SEDD 的密度比参数化与 score entropy 训练目标。

39. Ou 等 (2024) . [Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data](#).

阅读重点：吸收扩散的 score 与干净数据条件分布之间的关系，以及由此得到的重参数化。

40. Ren、Chen、Zhu、Guo、Chen、Rotskoff、Tao、Ying (2025) . [Fast Solvers for Discrete Diffusion Models: Theory and Applications of High Order Algorithms](#).

阅读重点：高阶离散扩散求解器的构造、误差分析，以及相同计算预算下的效果。

41. Cho、Wu (2026) . [Minimax Optimality of Score Entropy Discrete Diffusion](#).

阅读重点：离散 score 估计的统计下界、接近匹配的上界，以及向生成分布误差的传递。

42. Jeon、Vuong、Tao (2026) . [VGB for Masked Diffusion Model: Efficient Test time Scaling for Reward Satisfaction and Sample Editing](#).

阅读重点：将验证器引导的回溯从前缀树推广到掩码状态图，通过解掩码与重新掩码实现奖励引导生成和样本修复，并分析验证器误差下的复杂度保证。

讨论问题：统计样本复杂度与采样计算复杂度分别由什么决定？比较求解器时，时间步数、网络调用数和实际运行时间是否给出相同判断？

推理侧引导与粒子方法

阅读路线：条件采样与 Doob h 变换，Feynman Kac 粒子校正，再到漂移控制和路径空间 SMC。

研究如何通过引导、重加权与重采样，将预训练扩散模型用于条件生成和逆问题。以 Doob h 变换、Feynman Kac 公式和随机控制为主要工具，分析目标分布、计算预算与近似误差之间的关系。

43. Wu、Trippe、Naesseth、Blei、Cunningham (2023) . [Practical and Asymptotically Exact Conditional Sampling in Diffusion Models](#)。

基础阅读：Twisted Diffusion Sampler 如何把启发式条件引导嵌入 SMC，并用权重修正保持渐进正确性；明确粒子数与时间离散化的不同极限。

44. Zhu、Ye、Liu、Wang、Chen (2026) . [Training Free Adaptation of Diffusion Models via Doob's \$h\$ Transform](#)。

阅读重点：Doob h 函数如何描述目标分布的变化，以及如何通过模拟近似相应的引导项。

45. Skreta 等 (2025) . [Feynman Kac Correctors in Diffusion: Annealing, Guidance, and Product of Experts](#)。

阅读重点：从目标密度演化推导权重与粒子校正，理解退火、guidance 和专家乘积。

46. Chen、Ren、Min、Ying、Izzo (2025) . [Solving Inverse Problems via Diffusion Based Priors: An Approximation Free Ensemble Sampling Approach](#)。

阅读重点：逆问题中的后验演化与加权粒子算法，以及模型误差和有限粒子误差。

47. Ren、Gao、Ying、Rotskoff、Han (2025) . [DriftLite: Lightweight Drift Control for Inference Time Scaling of Diffusion Models](#)。

阅读重点：如何联合设计漂移和粒子权重，以减轻权重退化并提高推理效率。

48. Chenyang Wang、Weizhong Wang、Ren、Blanchet、Lu (2026) . [Simple Approximation and Derivative Free Inference Time Scaling for Diffusion Models via Sequential Monte Carlo on Path Measures](#)。

阅读重点：利用 Girsanov 变换构造路径空间的重要性权重，以无需导数的重采样实现扩散模型引导，并理解路径层面与粒子层面 SMC 的等价关系。

讨论问题：不同算法是否逼近同一个目标分布？引导函数的近似误差、时间离散化误差与有限粒子误差如何区分？比较时应计入 score、奖励及其导数的调用成本，并考察粒子相关性和权重退化。

三 语言模型

从训练过程中形成的表示与计算机制出发，研究语言模型如何利用上下文、执行推理、使用额外计算，以及通过后训练改善能力。将可控模型中的数学分析与语言模型的实验现象联系起来，区分表达能力、训练可学习性、因果机制和统计泛化。

49. 苏剑林（2021，科学空间博客）. [Transformer 升级之路：2、博采众长的旋转式位置编码](#)。

结构基础导读：理解 RoPE 如何用位置相关旋转，让注意力内积中的位置项依赖相对位移。结合后文的表达能力与机制分析，讨论位置编码如何进入模型假设，以及代数结构能为长度外推提供什么依据。

训练动力学与可解释性

阅读路线：平均场测度动力学，单指标模型的信息指数与 SGD 学习时间，再到 Transformer 的能力形成、循环计算、回路分析和因果解释。

这一专题关注训练学到了什么，以及特征和计算机制如何随训练形成。先从两层网络与单指标模型学习描述训练动力学的数学工具，再与 Transformer 的训练现象和机制解释对照。阅读时明确网络宽度、输入维数、训练时间和步长各自的缩放，以及结论适用的模型与数据假设。

平均场极限与特征学习

把神经元看作相互作用粒子，以参数经验测度描述宽网络的训练。重点讨论平均场 PDE、Wasserstein 梯度流、有限宽度涨落与随机梯度噪声，可与第一部分 Chizat、Bach 的最优输运与全局收敛工作对照。

50. Mei、Montanari、Nguyen（2018）. [A Mean Field View of the Landscape of Two Layers Neural Networks](#)。

理论主读：从两层网络 SGD 推导参数经验分布的非线性 PDE，理解平均场极限如何保留特征学习，以及收敛和泛化分析依赖的假设。

51. Rotskoff、Vanden Eijnden（2018）. [Trainability and Accuracy of Neural Networks: An Interacting Particle System Approach](#)。

理论主读：将神经元视为相互作用粒子，通过大数定律与中心极限定理刻画平均场动力学、有限宽度涨落和 SGD 噪声，并讨论宽度与训练误差的关系。

52. Sirignano、Spiliopoulos（2018）. [Mean Field Analysis of Neural Networks: A Law of Large Numbers](#)。

证明选读：在网络宽度与训练迭代数共同增长的尺度下，证明参数经验测度收敛到非线性 PDE；重点理解弱收敛、极限方程唯一性与混沌传播的证明路线。

讨论问题：平均场极限需要怎样的网络参数化与训练尺度？有限时间内的经验测度收敛，如何与长时间训练和全局最优性联系？宽度、训练时间与输入维数的极限是否可以交换？

单指标模型与信息指数

在 single index 模型中，目标由输入在未知方向上的投影生成。当总体风险仅依赖与真实方向的相关性时，information exponent 刻画其在零相关处首个非零下降项的阶数。由此分析 SGD 从随机初始化建立相关性所需的时间，并结合 Hermite 展开比较信号、梯度噪声与样本复杂度。

53. Ben Arous、Gheissari、Jagannath (2020) . [Online stochastic gradient descent on non convex losses from high dimensional inference](#)。

理论主读：通过信息指数刻画随机初始化附近的有效信号，分析在线 SGD 达到弱恢复的搜索阶段与随后的下降阶段，以及样本复杂度对信息指数的依赖。

54. Bietti、Bruna、Sanford、Song (2022) . [Learning Single Index Models with Shallow Neural Networks](#)。

理论主读：在投影方向与链接函数均未知时，浅层网络如何通过梯度流学习单指标结构；理解优化景观、网络参数化与统计保证之间的联系。

55. Damian、Nichani、Ge、Lee (2023) . [Smoothing the Landscape Boosts the Signal for SGD: Optimal Sample Complexity for Learning Single Index Models](#)。

算法与理论主读：平滑损失如何改善随机初始化附近的梯度信噪比，缩小普通在线 SGD 与相关统计查询下界之间的差距；比较信息指数、损失设计与样本复杂度。

56. Damian、Pillaud Vivien、Lee、Bruna (2024) . [Computational Statistical Gaps in Gaussian Single Index Models](#)。

进阶阅读：区分 information exponent 与 generative exponent，理解标签变换如何暴露更低阶的信号，以及统计查询和低次多项式框架下的计算与统计差距。

讨论问题：信息指数如何依赖链接函数与损失，何时等于相应函数的 Hermite rank？平滑、非线性变换和数据结构如何改变可利用的信号？比较复杂度时，应明确在线新样本与样本复用、已知与未知链接函数，以及结论针对的算法访问模型。

Transformer 的训练动力学与能力形成

57. Kumar、Bordelon、Gershman、Pehlevan (2023) . [Grokking as the Transition from Lazy to Rich Training Dynamics](#)。

机制主读：在可控模型中追踪从核回归式拟合到特征学习的训练过程，理解初始特征与目标的谱对齐、特征学习速度如何影响延迟泛化。

58. Zhang、Frei、Bartlett (2023) . [Trained Transformers Learn Linear Models In Context](#)。

理论主读：分析单层线性注意力在预训练中的梯度流，证明适当初始化下的全局收敛，并考察训练所得上下文学习能力对协变量分布变化的敏感性。

59. Nichani、Damian、Lee (2024) . [How Transformers Learn Causal Structure with Gradient Descent](#)。

理论主读：理解注意力矩阵的梯度如何编码词元间的互信息，并利用数据处理不等式分析简化双层 Transformer 学习潜在因果结构与归纳头的机制。

60. Singh、Chan、Moskowitz、Grant、Saxe、Hill (2023) . [The Transient Nature of Emergent In Context Learning in Transformers](#)。

实证补充：在合成任务中比较上下文学习与参数内学习的竞争，考察训练损失持续下降时 ICL 能力为何可能先出现再消失，以及正则化如何改变这一过程。

Looped Transformer 的循环计算与推理

沿深度重复使用同一个 Transformer 模块，让隐状态逐步更新，以研究参数共享、有效深度与迭代算法的关系。阅读路线从 Universal Transformer 的循环结构出发，经由可编程计算与上下文学习，进入训练收敛理论、隐空间推理和推理侧扩展。区分参数优化的训练时间与一次前向计算中的循环时间，并将这一专题与后面的上下文学习和推理侧扩展对照。

61. Dehghani、Gouws、Vinyals、Uszkoreit、Kaiser（2018）. [Universal Transformers](#)。

基础阅读：沿深度共享参数并反复更新序列表示，引入按位置自适应停止机制；理解循环结构的归纳偏置，以及算法任务中的长度泛化问题。

62. Giannou、Rajput、Sohn、Kangwook Lee、Jason D. Lee、Papailiopoulos（2023）. [Looped Transformers as Programmable Computers](#)。

表达能力主读：通过显式构造权重与输入指令，让固定层数的循环 Transformer 执行程序与迭代算法；关注计算状态和循环次数如何编码，以及该构造与训练可学习性的距离。

63. Yang、Lee、Nowak、Papailiopoulos（2023）. [Looped Transformers are Better at Learning Learning Algorithms](#)。

实证主读：在上下文数据拟合任务中，研究循环结构及其训练方法如何促进迭代学习行为；比较参数量、循环次数与预测误差，并考察是否学到了可重复执行的更新规则。

64. Gatmiry、Saunshi、Reddi、Jegelka、Kumar（2024）. [Can Looped Transformers Learn to Implement Multi step Gradient Descent for In context Learning?](#)。

训练理论主读：在文中的线性回归与线性循环 Transformer 设置下，总体损失的全局最优解实现多步预条件梯度下降；重点阅读梯度支配条件、梯度流收敛证明及其模型和数据假设。

65. Saunshi、Dikkala、Zhiyuan Li、Kumar、Reddi（2025）. [Reasoning with Latent Thoughts: On the Power of Looped Transformers](#)。

理论与实验主读：比较参数量与有效深度对推理的作用，阅读循环模型求解迭代任务和模拟思维链的构造；注意模拟所需的额外工作空间，并讨论推理表现与记忆、困惑度之间的关系。

66. Geiping、McLeish、Jain、Kirchenbauer、Singh 等（2025）. [Scaling up Test Time Compute with Latent Reasoning: A Recurrent Depth Approach](#)。

实证主读：通过循环模块在隐空间反复计算，将推理预算转化为更大的有效深度；关注训练时的循环次数采样、截断反向传播和推理时的自适应停止，以及增加循环次数的收益与成本。

讨论问题：共享权重何时形成可重复执行的算法，何时导致误差积累或退化？如何区分训练得到的算法、表达能力构造与实验现象？增加循环次数能否泛化到训练时未见的深度和任务长度？比较循环计算与显式思维链时，应同时说明参数量、总计算量、工作空间与延迟。

选题建议：训练理论可主讲 Gatmiry 等，并与 Yang 等的实验对照；推理与计算复杂度可主讲 Saunshi 等，并与 Geiping 等的模型实验对照。其余两篇用于介绍结构来源与表达能力基础，具体主读论文与陆一平讨论确定。

内部机制与因果解释

67. Elhage、Nanda、Olsson 等（2021 技术文章）. [A Mathematical Framework for Transformer Circuits](#)。

框架补充：从残差流、QK 与 OV 回路及注意力头的组合出发，建立分析 Transformer 内部计算机制的基本语言，为后续归纳头与回路分析做准备。

68. Olsson、Elhage、Nanda 等（2022）. [In Context Learning and Induction Heads](#)。

机制主读：跟踪归纳头在训练中的形成及其与上下文学习能力变化的关系，比较小型纯注意力模型中的因果证据和较大模型中的相关证据。

69. Nanda、Chan、Lieberum、Smith、Steinhardt（2023）. [Progress measures for grokking via mechanistic interpretability](#)。

机制主读：在模加法任务中识别 Fourier 回路，通过消融和连续进度指标分析记忆、回路形成与清理三个训练阶段，解释测试性能骤变背后的渐进机制。

70. Bricken、Templeton、Batson、Chen、Jermyn 等（2023 技术文章）. [Towards Monosemanticity: Decomposing Language Models With Dictionary Learning](#)。

特征可解释性补充：用稀疏自编码器分解多义神经元，结合特征激活干预讨论可解释特征、重构质量与模型行为之间的关系。

讨论问题：训练损失的平滑下降为何可能伴随能力的突变或消退？找到一个相关的注意力头或特征，距离证明它是任务机制还差什么？参数消融、激活替换和输入干预分别能提供什么证据？在简化任务中识别的机制能否迁移到更大的语言模型？

选题建议：可围绕平均场极限与有限宽度效应、信息指数与 SGD 搜索时间、grokking 的特征学习与回路形成、循环结构与迭代算法、归纳头的观察与训练理论，或稀疏特征与因果验证组织报告。下一专题进一步讨论已形成的机制具有怎样的计算能力和统计保证。

上下文学习与推理的理论

阅读路线：贝叶斯解释与可控回归实验，注意力实现学习算法，再到算法选择、思维链表达能力、多智能体交互与采样规模定律。

围绕两个相互联系的问题组织阅读：**Transformer** 如何利用上下文实现学习算法？增加思维链、搜索或多次采样，如何改变计算能力与成功概率？报告中应明确区分表达能力、训练可学习性和统计泛化保证。

71. Xie、Raghuathan、Liang、Ma（2021）. [An Explanation of In context Learning as Implicit Bayesian Inference](#)。

理论主读：在隐马尔可夫混合模型中，预训练文档的潜在概念如何使上下文学习表现为贝叶斯推断；关注分布失配的条件。

72. Garg、Tsipras、Liang、Valiant（2022）. [What Can Transformers Learn In Context? A Case Study of Simple Function Classes](#)。

实证基础：在线性函数、稀疏函数及其他可控任务中检验上下文学习，建立与标准统计估计器比较的实验框架。

- 73.** Akyürek、Schuermans、Andreas、Ma、Zhou（2022）. [What learning algorithm is in context learning? Investigations with linear models.](#)
理论与实验主读：Transformer 实现梯度下降与岭回归的构造，以及深度、噪声和宽度如何影响训练后模型对应的估计器。
- 74.** von Oswald 等（2022）. [Transformers learn in context by gradient descent.](#)
理论主读：线性自注意力与回归梯度下降的对应；区分显式权重构造、训练得到该机制的实验证据及一般 Transformer 的情况。
- 75.** 苏剑林（2025，科学空间博客）. [线性注意力简史：从模仿、创新到反哺.](#)
中文导读：从线性注意力的递推形式进入在线学习、TTT 与 DeltaNet，配合上下文梯度下降论文阅读。比较沿序列更新的状态、Looped Transformer 沿深度的循环，以及训练时的参数更新，明确各自的优化变量。
- 76.** Bai、Chen、Wang、Xiong、Mei（2023）. [Transformers as Statisticians: Provable In Context Learning with In Context Algorithm Selection.](#)
阅读重点：Song Mei 等关于上下文内实现统计学习算法、算法选择及相应统计保证的工作。
- 77.** Feng、Zhang、Gu、Ye、He、Wang（2023）. [Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective.](#)
阅读重点：算术、方程和动态规划任务中的 CoT 理论，关注问题构造与复杂度假设。
- 78.** Merrill、Sabharwal（2023）. [The Expressive Power of Transformers with Chain of Thought.](#)
阅读重点：William Merrill 等如何刻画思维链长度与计算表达能力之间的关系。
- 79.** Li、Liu、Zhou、Ma（2024）. [Chain of Thought Empowers Transformers to Solve Inherently Serial Problems.](#)
阅读重点：Zhiyuan Li 等如何从串行计算角度解释 CoT 的作用，以及深度、精度与长度的权衡。
- 80.** Chen、Raventós、Cheng、Ganguli、Druckmann（2025）. [Rethinking Fine Tuning when Scaling Test Time Compute: Limiting Confidence Improves Mathematical Reasoning.](#)
阅读重点：训练目标与 pass@N 的不一致，以及模型置信度如何影响多次采样的收益。
- 81.** Du、Li、Torralba、Tenenbaum、Mordatch（2023）. [Improving Factuality and Reasoning in Language Models through Multiagent Debate.](#)
实证主读：多轮辩论如何改变推理与事实准确率；比较独立投票、共享信息和反复交互，并控制总生成与验证预算。
- 82.** Halder、Banerjee、Pehlevan（2026）. [Jailbreak Scaling Laws for Large Language Models: Polynomial Exponential Crossover.](#)
阅读重点：不同攻击条件下，采样次数与成功率的关系，以及统计物理模型如何解释规模定律的变化。

83. El 等 (2026) . [Physics of Agents: Statistical Mechanics Predicts Collective Behavior of AI Agents](#).

阅读重点：智能体交互中的共识、极化与群体准确率，及其统计力学建模。

讨论问题：存在一个能执行算法的 Transformer，距离实际训练得到这个模型还有哪些问题？固定预算下，延长单条推理链、增加候选数和改进验证器分别解决什么困难？

推理侧扩展与验证

从 Best of N 出发，比较拒绝采样、过程验证器、概率回溯和序贯蒙特卡洛。Dylan J. Foster 及其合作者的工作提供覆盖率、奖励误差与计算复杂度的理论分析；近似奖励模型的理论进一步刻画序列生成中的误差与计算效率。Dean Foster 参与的《Mind the Gap》将在后训练专题中连接自我验证与能力改进。

84. Maddison、Tarlow、Minka (2014) . [A* Sampling](#).

理论补充：Gumbel 过程如何将采样转化为随机优化，以及上界驱动搜索怎样得到正确样本；讨论该方法所需的界在语言模型中是否可得。

85. Snell、Lee、Xu、Kumar (2024) . [Scaling LLM Test Time Compute Optimally can be More Effective than Scaling Model Parameters](#).

背景阅读：问题难度如何影响搜索和答案修正的收益，以及固定预算下的计算分配；区分实验中的最优配置与理论最优性。

86. Lightman 等 (2023) . [Let's Verify Step by Step](#).

实证基础：过程监督与结果监督如何影响数学推理中的验证器和候选选择；关注人工标注成本、主动学习与评价设置。

87. Huang、Block、Liu、Jiang、Krishnamurthy、Dylan J. Foster (2025) . [Is Best of N the Best of Them? Coverage, Scaling, and Optimality in Inference Time Alignment](#).

阅读重点：覆盖率与奖励模型误差如何限制 Best of N；拒绝采样中的悲观原则，及其在论文条件下避免随采样数增加而退化的保证。

88. Rohatgi、Shetty、Saless、Li、Moitra、Risteski、Dylan J. Foster (2025) . [Taming Imperfect Process Verifiers: A Sampling Perspective on Backtracking](#).

阅读重点：过程验证器误差为何会在生成中被放大，以及树上随机游走和概率回溯如何改善对误差的鲁棒性。

89. Zhao、Brekelmans、Makhzani、Grosse (2024) . [Probabilistic Inference in Language Models via Twisted Sequential Monte Carlo](#).

理论与算法主读：twist 函数如何估计未来价值并引导粒子采样，及其与软强化学习、配分函数估计和分布误差评价的联系。

90. Golowich、Chen、Rohatgi、Singhal、Domingo Enrich、Dylan J. Foster、Krishnamurthy (2026) . [Reject, Resample, Repeat: Understanding Parallel Reasoning in Language Model Inference](#)。

阅读重点：从粒子滤波分析并行推理中的筛选与重采样，比较 SMC 的非渐近误差、算法改进与下界；特别注意采样误差与最终答案准确率的区别。

91. Zhu、Lu (2026) . [On the Power of \(Approximate\) Reward Models for Inference Time Scaling](#)。

理论主读：以近似奖励模型的 Bellman 误差刻画 SMC 推理的效率，分析误差随推理长度缩小时，计算复杂度从指数到多项式改善的条件。

讨论问题：先明确目标是逼近某个分布、提高 pass@N，还是提高最终选中答案的准确率。比较时应计入生成、验证与搜索的成本；分析验证器误差是否随预算增加被放大，以及有限粒子数、样本相关性和权重退化的影响。

后训练的理论

阅读路线：InstructGPT 的实际流程，RLHF、DPO 与 IPO，再选择覆盖率和自我改进、在线探索或 RLVR 课程学习深入。

研究预训练之后的监督微调、偏好优化、强化学习与自我改进。核心问题是：已有知识和基础模型分布如何影响后训练效率？推理侧搜索的收益怎样通过训练转化为单次生成的能力？在线探索和课程学习又能在什么条件下改善覆盖率与泛化？

Dean Foster 的《Mind the Gap》和 Dylan J. Foster 及其合作者关于 sharpening、coverage 与 exploration 的系列工作，可作为联系推理侧计算与后训练的阅读主线。以下按偏好学习、自我改进、探索与推理泛化三个方向组织，选题时可各取一篇主论文，再与相关工作对照。

偏好学习 对齐与过度优化

92. Ouyang 等 (2022) . [Training language models to follow instructions with human feedback](#)。

实证基础：InstructGPT 中 SFT、奖励模型与策略优化的完整流程，为后续 RLHF 理论提供实际问题 and 评价背景。

93. Zhu、Jordan、Jiao (2023) . [Principled Reinforcement Learning with Human Feedback from Pairwise or K wise Comparisons](#)。

基础阅读：在线性奖励及偏好比较模型下，奖励估计误差如何传递到策略质量，以及悲观估计和数据覆盖假设的作用。

94. Rafailov、Sharma、Mitchell、Ermon、Manning、Finn (2023) . [Direct Preference Optimization: Your Language Model is Secretly a Reward Model](#)。

基础阅读：KL 正则化下奖励与最优策略的关系，以及 DPO 损失的推导；讨论总体目标的等价性与有限样本、模型失配之间的差别。

95. Azar、Rowland、Piot、Guo、Calandriello、Valko、Munos (2023) . [A General Theoretical Paradigm to Understand Learning from Human Preferences](#)。

理论主读： Ψ PO 的统一偏好目标与 IPO，分析从成对偏好到标量奖励的建模假设，以及 RLHF 和 DPO 的潜在问题。

96. Huang、Zhan、Xie、Lee、Sun、Krishnamurthy、Dylan J. Foster (2024) . [Correcting the Mythos of KL Regularization: Direct Alignment without Overoptimization via Chi Squared Preference Optimization](#)。

阅读重点：KL 正则化为何可能无法控制离线对齐中的过度优化，以及 χ^2 正则化、悲观原则与单策略覆盖条件之间的联系。

讨论问题：偏好预测准确是否足以保证策略好？从训练分布移向高奖励区域时，估计误差如何被放大？DPO 与 χ PO 的保证分别依赖什么偏好模型、函数类与覆盖条件？

预训练覆盖与自我改进

97. Song、Zhang、Eisenach、Kakade、Dean Foster、Ghai (2024) . [Mind the Gap: Examining the Self Improvement Capabilities of Large Language Models](#)。

阅读重点：生成与验证能力之间的差距如何决定自我改进空间，以及自我验证、筛选或重加权、蒸馏之间的联系；区分数学条件与关于预训练规模的实验观察。

98. Huang、Block、Dylan J. Foster、Rohatgi、Zhang、Simchowitz、Ash、Krishnamurthy (2024) . [Self Improvement in Language Models: The Sharpening Mechanism](#)。

阅读重点：把高质量序列的概率质量提高理解为 sharpening，分析 SFT 与 RLHF 自我改进的样本复杂度，以及在线探索怎样改变对初始覆盖率的要求。

99. Chen、Huang、Golowich、Malladi、Block、Ash、Krishnamurthy、Dylan J. Foster (2025) . [The Coverage Principle: How Pre Training Enables Post Training](#)。

阅读重点：下一词预测如何促成对优质回答的覆盖，覆盖率为何可能比交叉熵更好地预测后训练与 Best of N 的效果，以及相应泛化界的假设和复杂度依赖。

讨论问题：如何定义并估计覆盖率和生成验证差距？把多次采样得到的好答案蒸馏回模型，节省了什么成本？由单个问题上的分布改善推到新问题上的泛化，还需要哪些假设？

在线探索 课程学习与推理泛化

100. Rafailov、Hejna、Park、Finn (2024) . [From r to Q*: Your Language Model is Secretly a Q Function](#)。

理论主读：把序列级 DPO 放回词元级 MDP，理解隐式 Q 函数、Bellman 关系、信用分配与推理侧搜索之间的联系。

101. Xie、Dylan J. Foster、Krishnamurthy、Rosset、Awadallah、Rakhlin (2024) . [Exploratory Preference Optimization: Harnessing Implicit Q* Approximation for Sample Efficient RLHF](#)。

阅读重点：DPO 与 Q^* 逼近的联系，XPO 的探索奖励如何改善在线偏好学习，以及样本效率结论所需的探索条件。

102. Dylan J. Foster、Mhammedi、Rohatgi (2025) . [Is a Good Foundation Necessary for Efficient Reinforcement Learning? The Computational Role of the Base Model in Exploration](#)。

阅读重点：在采样 oracle 与线性 softmax 框架下，覆盖率对运行时间和样本复杂度的不同作用，以及 SpannerSampling 与多轮探索的计算收益。

103. Rajaraman、Huang、Dudik、Schapire、Dylan J. Foster、Krishnamurthy (2026) . [Learning to Reason with Curriculum I: Provable Benefits of Autocurriculum](#)。

阅读重点：根据当前能力自适应选择训练问题，怎样降低 SFT 的示范需求与 RL 的计算成本；比较数据选择带来的收益及访问教师或奖励的条件。

104. Rajaraman、Huang、Dudik、Schapire、Dylan J. Foster、Krishnamurthy (2026) . [Learning to Reason with Curriculum II: Compositional Generalization](#)。

进阶阅读：在半自动机的序列计算模型中，分解与组合子问题如何降低监督量，并将 RLVR 对整条序列的覆盖要求转化为较短子块上的要求。

105. DeepSeek AI 等 (2025) . [DeepSeek R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#)。

实证基础：比较直接强化学习、冷启动、多阶段训练与蒸馏，为 RLVR 理论选择需要解释的现象；不把基准成绩当作普遍的能力扩展定理。

106. Wen 等 (2025) . [Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs](#)。

补充阅读：结果奖励在何种简化假设下会激励正确推理，GRPO 的训练动力学，以及 pass@K 与同时检查中间步骤的 CoT pass@K 如何给出不同评价。

讨论问题：奖励可验证解决了什么，又留下了哪些探索与泛化困难？样本复杂度小是否意味着实际计算高效？课程学习改善的是数据选择、信用分配还是组合泛化？应分别说明定理分析的任务、训练样本上的改进与实验中的跨题泛化。

选题建议：可从 DPO 与 χ PO、Mind the Gap 与 Sharpening、Coverage Principle 与 Best of N、XPO 与基础模型探索、Autocurriculum I 与 II 中任选一组。每组围绕一个问题展开，并结合个人兴趣继续寻找相关论文，与陆一平讨论主讲范围。

四 神经网络与数据的规模定律

阅读路线：经验规模定律与计算分配，核谱和可解模型，再分别研究网络的随机极限与数据统计结构。

模型规模 计算预算与随机极限

先理解经验规模定律如何支持资源配置，再研究宽度、深度和参数化的极限如何产生可分析的数学模型。阅读时区分两类结论：关于性能与资源的规模定律，以及关于网络随机过程的缩放极限。

107. Hestness 等 (2017) . [Deep Learning Scaling is Predictable, Empirically](#)。

基础阅读：较早系统研究不同任务中数据量、模型规模与泛化误差的关系，理解如何用小规模实验预测更大规模的表现。

108. Kaplan 等（2020）. [Scaling Laws for Neural Language Models](#)。

背景阅读：模型参数、训练数据与计算量如何共同影响语言模型损失。

109. Hoffmann 等（2022）. [Training Compute Optimal Large Language Models](#)。

背景阅读：在固定计算预算下，如何分配模型规模与训练数据量。

110. Bahri、Dyer、Kaplan、Lee、Sharma（2021）. [Explaining Neural Scaling Laws](#)。

理论主读：区分方差受限与分辨率受限的规模区间，将核谱、数据流形和模型规模联系起来；比较不同机制给出的指数。

111. Maloney、Roberts、Sully（2022）. [A Solvable Model of Neural Scaling Laws](#)。

进阶阅读：在可解的数据与随机特征模型中推导数据量和参数量的联合规模定律，分析谱幂律与性能饱和。

112. Mufan Li（2026 讲义）. [Lecture Notes on Neural Network Scaling Limits](#)。

阅读建议：第 2 章核极限、第 4 章特征学习与平均场；重点选读第 5、6 章的深度与宽度缩放及 SDE 极限。

113. Zhang、Morwani、Vyas、Wu、Zou、Ghai、Dean Foster、Kakade（2024）. [How Does Critical Batch Size Scale in Pre training?](#)。

阅读重点：临界批大小如何随数据量与模型规模变化，以及无限宽网络和最小二乘模型如何解释实验规律；区分计算效率与实际训练时间。

114. Cont、Rossier、Xu（2022）. [Asymptotic Analysis of Deep Residual Networks](#)。

理论主读：训练后权重的不同缩放如何使深层残差网络趋向 ODE、SDE 或其他极限，并分析隐藏状态与反向传播动力学的对应关系。

数据统计结构 自然语言熵与规模定律

研究数据量之外的因素如何影响学习难度，包括长程相关、条件熵、下一词分布的歧义和合成数据的尾部结构。重点讨论：能否从数据的可测统计量预测规模定律，以及如何把“数据质量”转化为明确的数学对象。

115. Shannon（1951）. [Prediction and Entropy of Printed English](#)。

基础阅读：通过预测任务估计语言熵与冗余，以及上下文长度的作用。

116. Scheibner、Smith、Bialek（2025）. [Large language models and the entropy of English](#)。

阅读重点：利用语言模型研究条件码长、长程依赖及其随上下文长度的变化。

117. Cagnetta、Raventós、Ganguli、Wyart（2026）. [Deriving Neural Scaling Laws from the statistics of natural language](#)。

阅读重点：由词元相关性和条件熵的衰减，预测数据受限情形下的规模定律指数。

118. Zucchet、Lee、Linderman（2026）. [Language models suffer from a curse of ambiguity](#)。

阅读重点：下一词分布的歧义如何影响表示容量、拟合难度和采样噪声。

119. Li 等（2024）. [DataComp LM: In search of the next generation of training sets for language models](#)。

实证补充：在控制训练流程与计算预算的框架中比较数据筛选和去重，讨论数据质量改变的是规模定律的系数、指数还是适用区间。

120. Dohmatob、Feng、Yang、Charton、Kempe（2024）. [A Tale of Tails: Model Collapse as a Change of Scaling Laws](#)。

阅读重点：合成数据改变尾部统计后，规模定律及模型能力如何随代际训练发生变化。

讨论问题：模型交叉熵、数据条件熵与熵率如何区分？高熵是否意味着低质量？跨语料或分词方案比较时，应如何统一测量单位与上下文？

五 神经网络优化器

阅读路线：Adafactor 与矩阵预条件，再到 SOAP、Muon、优化几何、隐式偏置与宽度缩放。

从矩阵预条件、非欧几里得几何和网络参数化出发，理解 Shampoo、SOAP、Muon 等优化器。重点比较它们利用的梯度结构、每步代价、收敛条件与实际训练表现，并讨论学习率如何随网络宽度变化及在不同规模间迁移。

121. Shazeer、Stern（2018）. [Adafactor: Adaptive Learning Rates with Sublinear Memory Cost](#)。

基础阅读：通过行列统计压缩二阶矩存储，并分析更新裁剪和参数相对尺度；与 Shampoo、SOAP 的矩阵结构利用方式对照。

122. Gupta、Koren、Singer（2018）. [Shampoo: Preconditioned Stochastic Tensor Optimization](#)。

基础阅读：保留张量结构的预条件方法，以及相应的计算、存储与收敛分析。

123. Vyas 等（2024）. [SOAP: Improving and Stabilizing Shampoo using Adam](#)。

阅读重点：在预条件矩阵特征基中使用 Adam 的思路，以及预条件更新频率的影响。

124. Keller Jordan（2024 技术文章）. [Muon: An optimizer for hidden layers in neural networks](#)。

基础阅读：Muon 的矩阵更新、正交化近似及原始设计动机。

125. 苏剑林（2024，科学空间博客）. [Muon 优化器赏析：从向量到矩阵的本质跨越](#)。

中文导读：从谱范数约束下的最速下降方向理解 Muon，联系 SVD、Newton Schulz 近似与 Shampoo。配合 Muon 原始文章及两种优化几何论文，讨论方向、步长、动量和矩阵分块如何共同决定实际更新。

126. Liu 等（2025）. [Muon is Scalable for LLM Training](#)。

阅读重点：Muon 扩展到大模型训练时的权重衰减、更新尺度与实际计算效率。

127. Soudry、Hoffer、Nacson、Gunasekar、Srebro（2017）. [The Implicit Bias of Gradient Descent on Separable Data](#)。

理论主读：在线性可分数据与相应损失条件下，梯度下降为何趋向最大间隔方向；联系优化几何、隐式正则化与泛化。

128. Xie、Wang、Wu、Li（2025）. [A Tale of Two Geometries: Adaptive Optimizers and Non Euclidean Descent](#)。

阅读重点：自适应优化器与最速下降分别依赖的光滑性和随机方差条件。

129. Yang 等（2022）. [Tensor Programs V: Tuning Large Neural Networks via Zero Shot Hyperparameter Transfer](#)。

补充阅读： μP 参数化与从小模型向大模型迁移超参数的基本思路。

130. 苏剑林（2025，科学空间博客）. [高阶 MuP：更简明但更高明的谱条件缩放](#)。

中文导读：以权重和更新量的谱范数理解特征尺度、学习率与宽度的关系，配合 Tensor Programs V 和矩阵算子范数论文阅读。推导时标出梯度有效低秩、Adam 近似及尺度估计的条件，并联系第四部分的宽度极限。

131. Xu、Li、Lu（2026）. [On the Width Scaling of Neural Optimizers Under Matrix Operator Norms I: Row/Column Normalization and Hyperparameter Transfer](#)。

阅读重点：矩阵算子范数、行列归一化和网络复合结构如何联系宽度缩放与学习率迁移；可邀请陆一平介绍。

讨论问题：比较优化器时，是否控制了计算预算、每步开销和调参成本？宽度无关的理论界、实验中的稳定学习率和最优学习率迁移之间分别是什么关系？